Regression metamodel summarization of model behaviour

Kleijnen, J.P.C.

Publication date:
1982

FEW

faculteit der economische wetenschappen

RESEARCH MEMORANDUM

[redacted]

TILBURG UNIVERSITY
DEPARTMENT OF ECONOMICS
Postbus 90153 - 5000 LE Tilburg
Netherlands

Foto

FEW
120

REGRESSION METAMODEL SUMMARIZATION OF MODEL BEHAVIOUR

J.P.C. Kleijnen
Katholieke Hogeschool Tilburg
5000 LE Tilburg, Netherlands

A contribution to:

Encyclopedia of Systems and Control,

Editor - in - Chief: M.G. Singh

Subject Editor: B.P. Zeigler

Publisher: Pergamon, Press, Oxford.

November 1982.

Regression Metamodel Summerization of Model Behaviour

This contribution surveys regression analysis of both deterministic and stochastic simulation models. Regression analysis is a statistical technique summarizing how the simulation model's output reacts to changes in the model's input. The technique should be accessible for the practitioner with a knowledge of basic statistical theory, e.g. he should be familiar with the t test. [A few comments meant for the specialist are placed within square brackets.]

Summarization of model behaviour is necessary for many reasons. Although simulation is an extremely flexible technique used to study many practical problems, the technique has one major drawback: strictly speaking its results are valid only for the specific parameter values and mathematical relationships of the executed simulation program. For example, if a queuing problem is studied and 100,000 customers are simulated then their average waiting time is a valid estimator conditional on specific parameter values (such as the arrival intensity λ and the service rate μ) and the specific queuing discipline (e.g., first-come-first-served, FCFS). To remedy this limitation the simulation analyst runs his program for several combinations of parameter values. (Besides changing quantitative parameters, the analyst may make qualitative changes, e.g., he may change the queuing discipline from FCFS to shortest-jobs-first; although regression analysis applies to both types of changes, this contribution focusses on quantitative changes; see KLeijnen (1983a) for qualitative variables.) Executing the program for many parameters combinations is necessary for sensitivity and robustness analysis, validation, optimization, what - if questions, etc. (these concepts are discussed in other contributions). Regression analysis aims at interpreting the great mass of data produced by the typical simulation program. Reams of paper output and numerous tables with numbers do not provide insight!

1. Regression Metamodeling: Common Sense Formalized and Extended

Suppose a collection of data has become available, i.e., the simulation program has been run for several combinations of parameter values and output has been recorded. (Kleijnen, 1983a, discusses efficient and effective techniques for specifying the parameter combinations actually run.) It is common sense to analyze the set of experimental data by plotting the input/output combinations. The analyst fits a curve by hand, and he concludes whether a parameter has an important effect on the output. Note that sometimes data are presented in tabular form. However, the analyst recognises patterns much easier when graphical representation is used.

The common sense approach is formalized and extended by regression analysis. (At the same time common sense and theoretical knowledge are the basis of the regression model; see Sect. 3.) Fitting the curve by hand is replaced by determining a curve that fits best tot the observed output data where "best" may be formalized as "minimum sum of squared deviations", briefly called Least Squares. (Other criteria than Least Squares will be discussed in Sect. 3.) In the common sense approach judging the importance of a simulation parameter is subjective, e.g., changing the scale of the plot can blow up the importance. Regression analysis uses objective tests of significance like the t test. Further, if the analyst fits (say) a straight line, statistical methods can check the validity of the fitted curve. Finally, if there are many parameters in the simulation model, then it becomes cumbersome to make plots for each parameter. It becomes virtually impossible to detect interactions among these parameters. (Interactions are also discussed in Kleijnen, 1983a.) Regression extends the analysis into multiple dimensions. The necessary formulas will be presented in Sect. 2.

The regression model (or its common sense equivalent: the input/output plots) may be called a <u>metamodel</u>. This concept fits in a hierarchical modeling approach, i.e., from the "mess" of reality the analyst proceeds to a well-structured simulation model, and the relationship between the input and output of this simulation model is in turn modeled by the regression model. In symbols, in reality the output variable y is determined in an unknown way (f_0) by numerous variables v:

$$y = f_0(v_1, v_2, ...) \tag{1}$$

The simulation model (f_1) concentrates on relatively few variables or "parameters", say z_1 through z_k. Random simulation explicitly recognizes the approximate character of the representation and introduces an explicit stochastic component through the random number stream $\underset{\sim}{R}$ (bold-face italics denote matrices including vectors; capitals denote random variables; e.g.,

$P(\underset{\sim}{R}) = P(R_1)\ P(R_2)\ldots$ and $P(R_1 < r) = r$). Because deterministic simulation is just a special case of random simulation - $P(Z = z) = 1$ - the following representation results:

$$Y = f_1(z_1, z_2, \ldots, z_k, \underset{\sim}{R}) \tag{2}$$

Although the simulation program (f_1) generates a time series, this series is usually summarized by a few measures such as the average. This contribution concentrates on a single measure per simulation run. However, the methodology can also be applied to each individual measure charaterizing the time series of an output variable. A mathematical foundation for approximating f_1 in Eqn. (2) would be available if all variables z were continuous and f_1 were a nicely behaving function and $\underset{\sim}{R}$ would vanish. Then the Taylor series expansion would yield the first-order approximation:

$$y = \gamma_0 + \gamma_1\ z_1 + \ldots + \gamma_k\ z_k \tag{3}$$

In practice, however, the simulation model f_1 is more complicated. (Because the simulation model is so complicated the need for summarizing model behavior arises!) Therefore the analyst first specifies a metamodel like Eqn. (3) and next he tests the validity of that model; see Sect. 3. The following section discusses regression analysis in more detail.

2. Regression Analysis: Simple Statistical Formulas

Let the available collection of simulation data be specified by Table 1. There are n observations or simulation runs. There are q independent variables x, including the dummy variable x_0 equal to one: $x_{i0} = 1$ for $i = 1, \ldots, n$. These variables might correspond to the simulation parameters z in Eqn. (2) and (3): $x_1 = z_1, \ldots, z_{q-1} = z_k$. However, the independent variable x may also be a function of the simulation parameter z, e.g., $x = z^2$ or $x = \log(z)$; see Sect. 3. Moreover, the independent variable x may be a binary variable, e.g., $x_{i1} = 1$ if in simulation run i the qualitative factor queuing discipline equals FCFS, and $x_{i1} = 0$ if the queuing discipline equals shortest-jobs-first; see Kleijnen (1983a). Each run yields one observation y on the output (dependent variable, response). If there are multiple output variables or measures the analysis may be applied per dependent variable. [Multivariate analysis is also possible but mote difficult.] This output may be deterministic or random, although the textbooks discuss only random outputs.

Least Squares is a mathematical, not a statistical, problem formulation: given the observations y_i and a family of curves with parameters γ, $\hat{y} = f(x_0, \ldots, x_{q-1}; \gamma_0, \ldots, \gamma_{q-1})$ determine the parameter values $\hat{\gamma}$ which minimize $\sum_i (y-\hat{y}_i)^2$. If the model $\hat{y}$ is linear in the regression parameters γ, then the Least Squares solution is simple:

$$\hat{\underset{\sim}{\gamma}} = (\underset{\sim}{x}'.\underset{\sim}{x})^{-1}.\underset{\sim}{x}'.\underset{\sim}{y} \tag{4}$$

where $\underset{\sim}{x}$ denotes the n×q matrix of independent variables and $\underset{\sim}{y}$ denotes the vector of n simulation responses. The estimation of γ becomes a statistical problem if we assume that one of the following situations holds:

(i) Random simulation: The observed simulation response y is a realization corresponding to a specific random number stream $\underset{\sim}{r}$; see Eqn. (2). In other words, given the simulation parameters z or the independent variables x, there is population of possible response values y. The simplest statistical model specifies that the random variable Y is normally distributed, with an expected value depending linearly on $\underset{\sim}{x}$, i.e.

$$E(\underset{\sim}{Y}|\underset{\sim}{x}) = \underset{\sim}{x}.\underset{\sim}{\gamma} \tag{5}$$

, and with a constant (conditional) variance:

$$var(Y|\underset{\sim}{x}) = \sigma^2 \tag{6}$$

Moreover, the n simulation responses are assumed to be independent (independent random number streams in the n runs), so that the covariance matrix of $\underset{\sim}{Y}$ equals

$$\underset{\sim}{\sigma} = \sigma^2.\underset{\sim}{I} \tag{7}$$

when the identy matrix is denoted by $\underset{\sim}{I}$ (though it is not a random matrix a capital letter is traditionally used.) In other words, the errors $\underset{\sim}{E}$ defined by

$$\underset{\sim}{Y} = \underset{\sim}{x}.\underset{\sim}{\gamma} + \underset{\sim}{E} \tag{8}$$

are normally and idependently distributed (NID) with zero mean and constant variance σ^2:

$$\underset{\sim}{E} \sim \text{NID}(\underset{\sim}{0}, \sigma^2.\underset{\sim}{I}) \tag{9}$$

(ii) Deterministic simulation: If the simulation responses are not random, then a regression model analogous to Eqn. (5) can still be formulated, and the deviations between this metamodel and the actual simulation responses may be assumed to satisfy Eqn. (9). (The n simulation responses combinations $\underset{\sim}{z}$ can be modeled as a sample from all possible combinations: these combinations result in the reponses $\underset{\sim}{y}$ with the corresponding errors $\underset{\sim}{e}$.)

Under the classical assumptions of Eqn. (9) it can be proved that the (mathematical) Least Squares algorithm leads to estimators of the regression parameters γ that are unbiased: $E(\hat{\underset{\sim}{\Gamma}}) = \underset{\sim}{\gamma}$ where $\hat{\underset{\sim}{\Gamma}}$ follows from Eqn. (4) if $\underset{\sim}{y}$ is replaced by $\underset{\sim}{Y}$. Note that $\hat{\underset{\sim}{\Gamma}}$ is a linear estimator, i.e., it is a linear transformation of the responses $\underset{\sim}{Y}$. Under the classical assumption the Least Squares estimator $\hat{\underset{\sim}{\Gamma}}$ is also the linear estimator with the smallest variance: Best Linear Unbiased Estimator (BLUE). These minimal variances can be derived from the following general statistical formula (which will be used again in Sect. 3). Define a random vector, say $\underset{\sim}{Y}_1$, and its linear transformation $\underset{\sim}{Y}_2$:

$$\underset{\sim}{Y}_2 = \underset{\sim}{a}.\underset{\sim}{Y}_1 \tag{10}$$

where $\underset{\sim}{a}$ is an $n_2 \times n_1$ matrix ($n_1 \geq 1$, $n_2 \geq 1$). Denote the covariance matrix of $\underset{\sim}{Y}_1$ by $\underset{\sim}{\sigma}_1$, where this $n_1 \times n_1$ matrix has elements

$$\sigma_{gg'} = E\{[Y_g - E(Y_g)].[Y_{g'} - E(Y_{g'})]\}$$

$$(g, g' = 1, \ldots, n_1) \tag{11}$$

Note that $g = g'$ yields $\sigma_{gg} = \sigma_g^2 = var(Y_g)$. It can be proved that the covariance matrix of $\underset{\sim}{Y}_2$ defined by Eqn. (10), equals

$$\underset{\sim}{\sigma}_2 = \underset{\sim}{a}.\underset{\sim}{\sigma}_1.\underset{\sim}{a}' \tag{12}$$

Applying Eqn. (12) to Eqn. (4) and (7) yields the covariance matrix of $\hat{\underset{\sim}{\Gamma}}$:

$$\underset{\sim}{\sigma}_\gamma = (\underset{\sim}{x}'.\underset{\sim}{x})^{-1}.\sigma^2 \tag{13}$$

The scalar σ^2 in Eqn. (13) was defined by Eqn. (6) and can be estimated through

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-q) \tag{14}$$

where q denotes the number of estimated regression parameters and $\hat{y}_i$ is the i-th component of

$$\hat{\underset{\sim}{y}} = \underset{\sim}{x}.\hat{\underset{\sim}{\gamma}} \tag{15}$$

For random simulation an estimator of σ^2 different from Eqn. (14), will be presented in Sect. 3. The main-diagonal elements of $\underset{\sim}{\sigma}_\gamma$

defined in Eqn. (13) are the variances of the regression parameter estimators, and their square roots s are the standard deviations or "standard errors". Testing whether γ_j equals zero - or more generally equals the value γ_j^0 - is based on the t statistic:

$$t = \frac{\hat{\gamma}_j - \gamma_j^0}{s_j} \qquad (j = 0, 1, \ldots, q-1) \tag{16}$$

The degrees of freedom of t equal the degrees of freedom of $\hat{\sigma}^2$, i.e., if Eqn. (14) is used then the degrees of freedom equal n-q. The above formulas can be found in many textbooks, e.g. (Draper and Smith 1966). Many statistical packages can be used to analyze the simulation data through regression modeling, e.g. SPSS.

3. Least Squares: Assumptions and Alternatives

The Least Squares estimators of Eqn. (4) are BLUE if a number of assumptions hold.

(i) The regression model is linear in its parameters γ.

This assumption does not mean that the metamodel is necessarily linear in the simulation parameters z introduced above Eqn. (2). For instance, the regression model may be a second degree polynomial in z_1:

$$Y = \gamma_0 + \gamma_1 . z_1 + \gamma_{11} . z_1^2 + E \tag{17}$$

so that the notation of Table 1 yields $x_{i0} = 1$, $x_{i1} = z_{i1}$ and $x_{i2} = z_{i1}^2$. A different example is:

$$Y = \gamma_0 + \gamma_1 . \log x + E \tag{18}$$

so that x_{i1} = log x. The last equation is equivalent to

$$Y^{*} = \gamma_0^{*} . x^{\gamma_1} . E^{*} \tag{19}$$

so that Y = log Y^{*}, γ_0 = log γ_0^{*} and E = log E^{*}. So a model not linear in its parameters can sometimes be transformed into a model which is linear in its parameters. However, if such a transformation cannot be found, nonlinear regression analysis has to be applied; see Bard (1974). In general, linear regression analysis is flexible enough for the summarization of simulation models. (Nonlinear regression is applied to, e.g., chemical experiments where enough theoretical knowledge is available to suggest a specific family of nonlinear models). The transformation of Eqn. (18) and (19) will be further discussed in (iv).

(ii) <u>The simulation responses Y have constant variances σ^2</u>. The regression model with the <u>additive</u> error term E in Eqn. (8) implies that Y and E have the same variance; see σ^2 in Eqn. (6) and (9). The assumption of a "homogenous" variance is unrealistic, except for deterministic simulation. If the random variable Y has an expected value that depends on $\underset{\sim}{x}$ - see Eqn. (5) - then it seems reasonable to assume that Y has a variance that also varies with $\underset{\sim}{x}$, i.e., Eqn. (6) is replaced by

$$\mathrm{var}(Y_i) = \sigma_i^2 \quad (i = 1, \ldots, n) \tag{20}$$

In random simulation each run should yield not only the point estimate y_i but also the standard error of y_i, denoted by $\hat{\sigma}_i$.

In the contribution by Schriber (1983) different techniques for estimating σ^2 are surveyed. For example, if y_i represents the average response of run i then the total run may be divided into b subruns (or batches), each yielding an average $\hat{y}_{ih}$ $(h = 1, \ldots, b)$ with var $(Y_i) = \mathrm{var}(Y_{ih})/b$ so that

$$\hat{\sigma}_i^2 = \sum_{h=1}^{b} (y_{ih} - y_i)^2 / \{(b-1).b\} \qquad (i = 1, \ldots, n) \tag{21}$$

Practice shows that these variance estimates $\hat{\sigma}_i^2$ differ greatly, say, by a factor 100 and more. So the assumption of a constant variance is in general unrealistic in random simulation. What is the alternative?

Intuitively, if a response has a high standard error, that response should receive less weight when fitting a curve. Formally, if the variances σ_i^2 were known, then the transformation

$$Y_i^* = \frac{Y_i}{\sigma_i} \qquad (i = 1, \ldots, n) \tag{22}$$

would result in constant variances: $\mathrm{var}(Y_i^*) = 1$. To the transformed output Y^* the Least Squares algorithm could be applied, resulting in <u>Weighted Least Squares</u>. In practice estimated variances $\hat{\sigma}_i^2$ are substituted into Eqn. (22) and the Estimated Weighted Least Squares estimators result:

$$\hat{\underset{\sim}{\gamma}} = (\underset{\sim}{x}' . \hat{\underset{\sim}{\sigma}}^{-1} . \underset{\sim}{x})^{-1} . \underset{\sim}{x}' . \hat{\underset{\sim}{\sigma}}^{-1} . \underset{\sim}{y} \tag{23}$$

where $\underset{\sim}{\sigma}$ is no longer given by Eqn. (7) but is now a diagonal matrix with main-diagonal elements given by Eqn. (21). If $\underset{\sim}{\sigma}$ were known then Eqn. (12) applied to Eqn. (23) would yield

$$\underset{\sim}{\sigma}_{\gamma} = (\underset{\sim}{x}' \cdot \underset{\sim}{\sigma}^{-1} \cdot \underset{\sim}{x})^{-1} \tag{24}$$

Simulation analysis showed that when $\underset{\sim}{\sigma}$ is estimated from at least five subruns (in Eqn. 21 $b \geq 5$) then Eqn. (24) is valid with $\underset{\sim}{\sigma}$ replaced by its estimate $\hat{\underset{\sim}{\sigma}}$; see (Kleijnen _et al_. 1981).

Note that (ordinary) Least Squares still yields unbiased estimators of the regression parameters γ but the variances of these estimators are no longer given by Eqn. (13). These variances can be found by applying Eqn. (12) to Eqn. (4). However, Estimated Weighted Least Squares results in more accurate estimators of γ so that important (significant) parameters are detected more frequently.

(iii) _The responses Y are independent_

In simulation experimentation the responses Y (and hence the errors E) can be forced to be independent through the use of independent random number streams, i.e., in Eqn (2) $\underset{\sim}{R}$ is sampled independently. However, sometimes variance reduction techniques such as common random numbers are used; see the contribution by Fishman (1983). Then the independence assumption is violated. If we can again estimate the covariance matrix of Y - now not a diagonal matrix - then Eqn. (23) and (24) apply, and the procedure is known as _Generalized Least Squares_; see (Draper and Smith 1966).

(iv) The responses Y are normally distributed.

The responses may be non-normal. For instance, if in Eqn. (19) Y^* were normal then the transformed response $Y = \log Y^*$ in Eqn. (18) would not be normal. (Usually Y is assumed to be normal and consequently Y^* is "lognormal".) If the responses are not normal then (Weighted) Least Squares still yields unbiased estimators and the standard errors are still given by Eqn. (24). However the t test of Eqn. (16) may be wrong. [If $\underset{\sim}{Y}$ is normal then the linear transformation $\hat{\underset{\sim}{\Gamma}}$ is normal, and the variance estimator S^2 is independent of Y.] To force the responses to be (more) normal, transformations may be applied. For instance, if the distribution of Y has a long tail at the right end, then log Y may be approximately normal. In simulation Y often denotes the average of the run. Limit theorems for independent and for autocorrelated variables explain why the average may be normally distributed. In long runs other responses than the mean are also approximately normal. So in practice nonnormality may be no serious problem.

Note that during the last decade many alternatives to (Weighted) Least Squares have been proposed. For instance, a different criterion may be: minimize the sum of relative absolute errors:

$$\text{Min} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i|} \tag{25}$$

Unfortunately the resulting estimators have unknown statistical properties, and the algorithm is more complicated. An alternative approach emphasizes that (Weighted) Least Squares is sensitive to outliers, and with nonnormal distributions such extreme responses may have a high chance of being observed. Several robust procedures have been proposed, and research on these procedures continues. Fortunately, in simulation it is much easier to check whether an observation is an outlier or is a truly representative observation: in random simulation the computer program can again be executed with one or more different random streams. In other fields than simulation it is often difficult to tell whether the suspicious observation is due to a measurement error, and in these fields true replication (i.e. observing the response for the same input conditions) is often difficult.

(v) <u>The regression model is correctly specified</u>.

If the simulation model in Eqn. (2) is adequately approximated by the regression metamodel, e.g. Eqn. (3), then the errors E have zero expectation; see Eqn. (9). Consequently the estimators of the regression parameters are unbiased, and the estimator of the variance σ^2 is also unbiased; see Eqn. (14). (The estimators $\hat{\sigma}_i^2$ in Eqn. (21) remain unbiased, even if the regression model is incorrect.) If the regression metamodel is not valid then the whole analysis breaks down. Fortunately there is a simple technique for testing the validity of the regression approximation. The technique is based on the following procedure (in the statistics literature different tests are proposed such as the lack-of-fit F test):

(i) Devise the model's general form.

(ii) "Calibrate" the model, i.e., specify its parameters.

(iii) Use the model tot forecast a "new" situation, i.e., a situation not used in steps (i) and (ii).

(iv) Compare the model's forecast to the actual response.

Such a procedure is generally used to validate models be it simulation models econometric models, or whatever type of model; also see the contribution by Sargent (1983). In the case of regression metamodeling the procedure becomes:

(i) Postulate a regression model, e.g. Eqn. (3). Moreover a statistical submodel must be specified; see the classical assumptions of Eqn. (9) or - more realistic in simulation - Eqn. (20).

(ii) From the sample of n observations - see Table 1 - estimate the regression parameters γ, using (Weighted) Least Squares; see Eqn. (4) or Eqn. (23).

(iii) Define a new situation $\underset{\sim}{x}'_{n+1} = (1, x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,q-1}) / \underset{\sim}{x}'_i$ $(i = 1, \ldots, n)$. And forecast the response:

$$\hat{y}_{n+1} = \underset{\sim}{x}'_{n+1} \cdot \hat{\underset{\sim}{\gamma}} \tag{26}$$

(iv) Simulate that new situation $\underset{\sim}{x}_{n+1}$ and obtain the simulation response y_{n+1}. Obviously the forecast $\hat{y}_{n+1}$ and the actual simulation y_{n+1} will not be exactly equal. Large deviations are acceptable if the statistical submodel (see i) specified large variability σ^2. Therefore compute the normalized deviation:

$$t = \frac{\hat{y}_{n+1} - y_{n+1}}{(\hat{\sigma}^2_{n+1} + \tilde{\sigma}^2_{n+1})^{\frac{1}{2}}} \tag{27}$$

where $\hat{\sigma}^2_{n+1} = \hat{\mathrm{var}}(Y_{n+1})$ follows from Eqn. (21), and $\tilde{\sigma}^2_{n+1} = \mathrm{var}(\hat{Y}_{n+1})$ follows from Eqn. (12) and (26):

$$\tilde{\sigma}^2_{n+1} = \underset{\sim}{x}'_{n+1} \cdot \hat{\underset{\sim}{\Omega}}_{\gamma} \cdot \underset{\sim}{x}_{n+1} \tag{28}$$

where $\hat{\sigma}_{\tilde{\gamma}}$ follows from Eqn. (13) or (24). The normalized deviation in Eqn. (27) may be compared to the critical values for the t statistic, or simpler, for the standard normal variable N(0,1).

Above a single validation run (namely run n+1) was used. There is a trick, however, to obtain many simulation runs for validation of the regression metamodel, provided $n > q$, i.e., there are more runs than regression parameters in Tabel 1. If $n > q$ then one run can be deleted (say sun 1) and the regression parameters can still be estimated from the remaining n-1 runs. The deleted run (run 1) can next be forecasted using Eqn. (26), and the forecast error can be computed using Eqn. (27). The trick continues as follows: Now a different run is deleted (say run 2 is deleted and run 1 is again added to the data available for estimation of the regression parameters γ), etc. This <u>permutation</u> or cross-validation approach yields n validation runs! Now a statistical complication arises: If n is, say, one hundred and the deviation in Eqn. (27) is tested with an α significance level of 5% (critical value 1.96) then five out of one hundred validation runs are expected to be false alarms (this follows from the definition of the α or type I error). To limit the false alarms to 5%, α is taken equal to 0.05/n (so-called Bonferroni approach); see Kleijnen (1975).

What are the alternatives if the regression metamodel is rejected?

(i) <u>Higher-order approximation</u>.

Above Eqn. (3) the Taylor series expansion was mentioned. If the first-order approximation of Eqn. (3) is rejected then crossproducts ($\gamma_{12}.z_1.z_2 + \ldots + \gamma_{k-1,k}.z_{k-1}.z_k$) can be added, representing interactions among the variables; see the contribution by Kleijnen (1983a). Also "pure" quadratic effect ($\gamma_{11}.z_1^2 + \ldots + \gamma_{kk}.z_k^2$) can be added, if the variables z are quantitative. If this second-order approximation is also rejected, then progressing to a third-order approximation is not recommended (difficult interpretation; many simulation runs required).

(ii) Transformations.

Before the very first form of the regression metamodel is postulated, the analyst should think hard about the "fundamental" variables in the simulation and hence in the regression model. For instance, in queuing problems theory shows - albeit for simplified analytical models - that the fundamental variable is not the arrival rate λ or the service rate μ, but their ratio $\rho = \lambda/\mu$, i.e., the traffic load. Consequently a metamodel with $x_1 = \rho$ is probably better than with $x_1 = \lambda$ and $x_2 = \mu$. If later on the regression model is rejected, the analyst may think again about the fundamental variables. But now he has more empirical data available (n simulation runs; regression results) which may guide his exploratory data analysis. Clearly statistical theory does not specify which variables should occur in the regression model!

4. Applications

The regression metamodel summarizes in an explicit form the relationship between the input x and the output y of the simulation model. This information can be used as follows.

(i) Sensitivity analysis, validation, what-if.

The simulation model may contain many parameters and relations. In a pilot investigation the model will be run for n different combinations of input. (Kleijnen (1983a) 's contribution shows how n can be kept small even if there is a great number of parameters.) If a regression parameter is significantly different from zero - see Eqn. (16) - then the corresponding simulation parameter (or qualitative variable) is important. If that important parameter reflects an environmental factor then the simulation model is valid only if sufficient data about the true value of that parameter is available. "What - if" questions will also concentrate on the important parameters.

(ii) Optimization and control.

Some regression parameters reflect variables that are under the user's control. The signs of the parameters show whether the corresponding decision variable should be increased or decreased. The relation magnitudes $\hat{\gamma}_j/\hat{\gamma}_{j'}$ show which variables are most critical. Actually, the regression metamodel specifies a response surface which can guide the search for an optimal - or at least a satisfactory - decision, using mathematical techniques such as Steepest Ascent; see (Montgomery and Bettencourt 1977). A "control" situation arises if the desirable output value is given and the corresponding input values are sought. The regression model immediately shows which combinations of input yield the desired output. Without the metamodel extensive trial-and-error is necessary.

Applications of regression metamodeling in simulation have started to appear. These applications concern simulations of job shops, steel plants, medical services, harbors, inventory control, etc.; see (Weeks and Fryer 1976), (Kleijnen et al. 1979), (Keijzer et al. 1982), Kleijnen (1983b).

Bibliography

Bard, Y., (1974), Nonlinear Parameter Estimation. Academic Press, Inc., New York.

Draper, N.R. and Smith, H., (1966), Applied Regression Analysis. John Wiley & Sons, Inc., New York.

Fishman, G.S., (1983), Variance Reduction Techniques in Stochastic Simulation. Encyclopedia of Systems and Control.

Keyzer, F., Kleijnen, J., Mullenders, E. and van Reeken, A., (1981) Optimization of priority class queues, with a computer center case study. American Journal of Mathematical and Management Sciences, 1, no. 4, pp. 341-358.

Kleijnen, J.P.C., (1974/1975), Statistical Techniques in Simulation. (In two volumes). Marcel Dekker, Inc., New York. (Russian translation: Publishing House "Statistics", Moscow, 1978).

Kleijnen, J.P.C., (1983), Design of (Stochastic) Simulation Experiments. Encyclopedia of Systems and Control.

Kleijnen, J.P.C., (1983), Statistical Tools for Simulation Practitioners. Marcel Dekker, Inc., New York.

Kleijnen, J.P.C., Brent, R. and Brouwers, R., (1981). Small-Sample Behavior of Weighted Least Squares in Experimental Design Applications. Communications in Statistics, Simulation and

Computation, B10, no. 3, pp. 303-313.

Kleijnen, J.P.C., van den Burg, A.J. and van der Ham, T., (1979). Generalization of Simulation Results: Practicality of Statistical Methods. European Journal of Operational Research, 3, pp. 50-64.

Montgomery, D.C. and Bettencourt, V.M., (1977), Multiple Response Surface Methods in Computer Simulation. Simulation, 29, no. 4, pp. 113-121.

Sargent, (1983), Validation of Models: General Approaches. Encyclopedia of Systems and Control.

Schriber, (1983), Statistical Analysis of Simulation Output. Encyclopedia of Systems and Control.

Weeks, J.K. and Fryer, J.S., (1976), A Simulation Study of Operating Policies in a Hypothetical Dual - Constrained Job Shop. Management Science, 22, no. 12, pp. 1362-1371.

J.P.C. Kleijnen

Catholic University Tilburg

Tilburg, Netherlands

Index: least squares, metamodel, optimization, sensitivity analysis, validation, what - if.

Tabel 1. Regression Analysis Data

Observation	Independent Variables 0	1	2	... j	... q-1	Dependent variable	(Standard error)
1	1	x_{11}	x_{12}	x_{1j}	$x_{1,q-1}$	y_1	(s_1)
2	1	x_{21}	x_{22}	x_{2j}	$x_{2,q-1}$	y_2	(s_2)
⋮							
i	1	x_{i1}	x_{i2}	x_{ij}	$x_{i,q-1}$	y_i	(s_i)
⋮							
n	1	x_{n1}	x_{n2}	x_{nj}	$x_{n,q-1}$	y_n	(s_n)

IN 1981 REEDS VERSCHENEN:

96 B.B. Van der Genugten
Central limit theorems for least squares estimators in linear regression models including lagged variables.

97 A.J.M. Swanenberg
Rationing and price dynamics in a simple market-game.

98 A.J. Hendriks
Theoretische en practische problemen van het distributieplanologisch onderzoek.
Referaten themadag RSA op 14 mei 1981.

99 A.J.J. Talman en L. Van der Heyden
Algorithms for the linear complementarity problem which allow an arbitrary starting point.

100 J.P.C. Kleijnen
Cross-validation using the t statistic.

101 J.P.C. Kleijnen
Statistical aspects of simulation: an updated survey. Version 1.

102 Dr. Wim G.H. van Hulst
On the concept of divergence in the theory of industrial organization.

103 H. Gremmen, T. van Bergen, J. Hotterbeekx
Waar liggen de Nederlandse comparatieve voordelen? Voorlopige versie.

104 A.L. Hempenius en P.G.H. Mulder
Multiple Failure Rates and Observations of Time Dependent Covariables. (Part 1: Theory.)

105 M.H.C. Paardekooper
A Newton-like method for error analysis. Applied to linear continuous systems and eigenproblems.

106 G. van der Laan en A.J.J. Talman
Note of the path following approach of equilibrium programming.

IN 1982 REEDS VERSCHENEN:

107 Aart J. de Zeeuw
Hierarchical decentralized optimal control in econometric policy models.

108 Arie Kapteyn en Tom Wansbeek
Identification in Factor Analysis.

109 G. van der Laan en A.J.J. Talman
Simplical Algorithms for finding Stationary Points, a unifying description

110 Pieter Boot
Ekonomiese betrekkingen tussen Oost en West Europa.

111 B.B. van der Genugten
The asymptotic behaviour of the estimated generalized least squares method in the linear regression model.

112 Jack P.C. Kleijnen / Anton J. van Reeken
Principles of computer charging in a university-like organization.

113 H. Tigelaar.
The informative sample size for dynamic multiple equation systems with moving average errors.

114 Drs. H.G. van Gemert, Dr. R.J. de Groof en Ir. A.J. Markink
Sektorstruktuur en Economische Ontwikkeling

115 P.H.M. Ruys
The tripolar model: a unifying approach to change.

116 Aart J. de Zeeuw
Policy solutions for Mini-Interplay, a Linked Model for two Common Market Countries.

IN 1982 REEDS VERSCHENEN (vervolg):

117 F.J.M. van Doorne en P.H.M. Ruys
Die Struktur einer Sprechhandlung in Habermas' Forschungsprogramm. Formale Analyse mit den Mitteln des tripolaren Modells.

118 A.J.J. Talman en G. van der Laan
Simplicial Approximation of Solutions to the Non-Linear Complementarity Problem

119 Pieter Boot
The East German Planning System Reconsidered